A New GIS-Based Tool for the Assessment of Environmental Equity and Death Rates Near Superfund Sites in the Urban Counties of Washington State

Richard Hoskins PhD, MPH*

Director, GIS and Spatial Epidemiology Unit, Office of Epidemiology, Washington State Department of Health, Olympia, WA

Abstract

All the geocoded Superfund (SF) and Toxics Release Inventory (TRI) sites from the Environmental Protection Agency's Landview III database in three urban counties in Washington State were developed into a geographic information system (GIS) coverage with circular buffers. Using a census block group base coverage, a spatial overlay was used to estimate population, various socioeconomic (SES) variables, and death rates from several causes. Age-stratified, age-adjusted death rates and standard mortality ratios were calculated and adjusted using empirical Bayesian smoothing with a prior distribution developed from the whole state and one from each block group's nearest neighbors. This was done to stabilize rates where the rate sample variance was high. Using the results of the buffer overlay, a profile was developed for all sites together. To facilitate comparison with other areas, a control group coverage was developed by building similar buffers around 25,000 random points inside the study counties. Points under water and in other areas not likely to have SF/TRI sites were excluded. Similar to a Monte Carlo simulation, control points were sampled and an empirical distribution developed for each variable for statistical testing. In the buffered regions, low income, status as a minority, limited education, high population density, and a high proportion of people over 65 were associated with SF/TRI sites. For causes of death, the death rate from cancer around SF/TRI sites was marginally statistically significant. After applying Bayesian smoothing to stabilize the rates, the differences became even less.

Keywords: Bayesian smoothing, Monte Carlo, environmental equity, Superfund, disease rates

Introduction

This paper describes a methodology for assessing the characteristics of neighborhoods located around areas that contain toxic waste or facilities that emit toxic waste. We are interested in determining if the socioeconomic characteristics of people living near these sites are different from in other neighborhoods, and whether they experience different rates of disease or death. Whether rates of disease have any causal relationship to these sites, as always, remains elusive.

In Washington State, public health practitioners need ways to determine the characteristics of populations around these sites that take geography into account. It is

* Richard Hoskins, GIS and Spatial Epidemiology Unit, Office of Epidemiology, Washington State Department of Health, 1102 Quince St., Olympia, WA 98504-7812 USA (p) 360-236-4270; (f) 360-236-4245; E-mail: reh0303@doh.wa.gov

important to consider the spatial context of neighborhoods with respect to characteristics such as low income or other indicators of low socioeconomic status that may predispose them to living close to toxic sites. In addition, it is important in the assessment of outcome measures such as disease or death rates to be able to deal with small area problems that plague disease rate calculations, such as an apparently high number of cases for a low population.

Our methods could be viewed as leaning toward violating the principle in epidemiology that cautions against using the "ecologic fallacy," whereby we assign characteristics to an individual based on the group to which they belong. We believe, however, that useful assessment methods must also avoid the "atomistic fallacy," which fails to consider the social, economic, and geographic context of individuals in a public health assessment.

This paper presents a means of assessment based on a Monte Carlo method of developing a statistical distribution that can be used for statistical testing. To determine disease rates, we adjusted rates in a neighborhood using an empirical Bayesian method to help account for small area problems. This work is in its initial stages and, therefore, is not presented as a completed effort.

Methods

Monte Carlo Method

For Snohomish, King, and Pierce counties in western Washington State, we developed a geographic information system (GIS) coverage of the 257 Superfund (SF) and Toxics Release Inventory (TRI) sites from the US Environmental Protection Agency (EPA) Landview III database (1). Information in this database allows for placing the sites at street-level accuracy. For this pilot study, we did not distinguish between the type of toxic contained or being emitted at a site, or a site's status as a SF or TRI site. Around each area, we constructed concentric rings of circular buffers of 0.25-, 0.5-, 1-, and 2-kilometer radii. We then developed a GIS coverage of US Census block groups with population attribute data for age, sex, and race, and with economic data from the 1990 census and projections for intercensus years from data provided by GeoLytics (2) and the Claritas Corporation (3). Using the spatial overlay operation in the GIS software, Maptitude (Calipter Corporation, Newton, MA), we were able to estimate the population and other characteristics from the block groups (or partial block groups) in each concentric buffer (4). Taking all the sites in one area, we were able to determine a profile of the residents within each of the buffer radii for population, income, education, and other census variables for the SF/TRI sites.

A control set was developed against which to compare this profile. In the three counties, we randomly assigned 25,000 points in the study area. The same size buffers were used and the same spatial overlays performed for each point as was done for the toxic sites. Excluded were points that fell in water areas (e.g., Puget Sound, Lake Washington) or in other locations where SF/TRI sites were not likely to be found. The distributions of these variables were developed by a Monte Carlo simulation. Using a uniform random number generator, repeated sets of control sites of 257 points each were drawn from the control group and a determination made of the value of the variable, such as population or income, using the results from the spatial overlays around each site. The result was a count histogram over the variable's range that indicated the likelihood that any particular value of a variable was found in the control set. A comparison of the toxic site profile with the control set was done by finding the value of the variable on the x-axis and computing the area from that point to the end of the curve. This results in a p-value for statistical comparison. An example showing the percentage of the population that is minority is presented in Figure 1, over which a Gaussian curve has been fit.

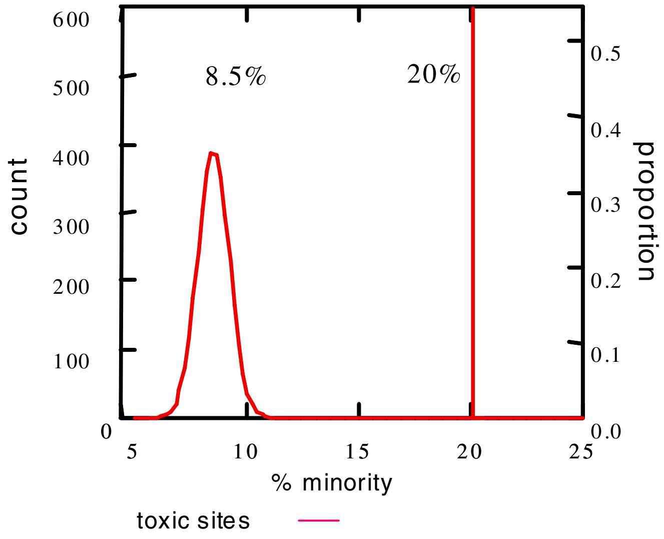

Figure 1 Percent minority race in a 1-kilometer buffer.

Empirical Bayesian Estimation of Death Rates

To determine the death rates from a variety of causes near the SF/TRI sites, and to compare them with the rates in the control areas, a spatial overlay procedure similar to the one described above was used. Death certificate data were geocoded to the street level and a point-in-polygon procedure was used to determine the number of deaths in each buffer area for a specific cause of death, age, race, and sex. This was done around each toxic area and control site. Using the buffer populations estimated by spatial overlay for the denominator, and point-in-polygon determinations for the numerator, the age-specific and age-adjusted rates were calculated for the various buffer sizes. These rates are likely to be unstable in areas where death counts are high in comparison with the underlying population. An adjustment can be made using an empirical Bayesian procedure; see Bailey and Gatrell (5) and Devine (6,7) for a good description of this method.

The adjustment of the age-specific and age-adjusted death rates is carried out according to the weighting scheme

$$\hat{r}_i^{Bayes} = \hat{w}_i r_i + (1 - \hat{w}_i)\hat{\gamma}$$

where

$$r_i = \text{observed rate for area } i$$

$$\hat{w}_i = \frac{\hat{\phi}}{(\hat{\phi} + \hat{\gamma} / n_i)} \quad \text{which is a shrinkage factor}$$

$\hat{\gamma}$ = mean death rate for the region (or nearest neighbors)

$\mathrm{n_i}$ = population in area i

and $\hat{\phi}$ is the weighted sample variance of the observed rates

$$\hat{\phi} = \frac{\sum \mathrm{n_i} (r_i - \hat{\gamma})^2}{\sum \mathrm{n_i}} - \hat{\gamma} / \overline{\mathrm{n}}$$

We adjusted the death rates in the various buffers according to this scheme, using the mean of the prior distribution for the whole state. In addition, we calculated the weighted sample variance using only the nearest neighbor block groups. Their contribution to the variance was further weighted depending on the intercentroid distance between the nearest neighbors. This yields a spatial or local Bayesian rate estimate, which was used to calculate the standard mortality ratio (SMR). Using this Bayesian rate or SMR rather than the observed rates around the toxic waste sites and control points addresses the high variability of the rate estimates and accounts for at least some of the spatial correlation.

Results

A comparison of several census variables for the toxic sites and several for the control group is shown in Table 1.

The differences between the values for all of the toxic sites compared with the controls (Table 1) were statistically significant at the 5% level or below, except for the variable percent college graduates living in the buffered areas around the toxic sites. Buffers at the other radii had similar results. Toxic sites were more like to have minority populations living in higher population densities. The sites are also more likely to have populations with less education, lower incomes, a lower percentage of children under 5 years of age, and a higher percentage of adults over 65 years of age.

The spatial or local Bayesian smoothing of rates for a variety of diseases was determined at the block group level throughout the three county area. Figure 2 shows the empirical distribution for SMRs for all cancer deaths between 1990 and 1996.

The SMR for all the control points was 112, which indicates a slightly higher cancer rate compared with the rest of the state. The SMR with no smoothing (i.e., the observed rate) for the toxic sites was 140, corresponding to a p-value of .12. The Bayesian and spatial Bayesian rates were 118 and 127, respectively, with p-values of .41 and .22. In this case, the Bayesian adjustments that took into account the rates of the entire state or

Table 1 Selected Census Variables for a One-Kilometer Buffer Comparing SF/TRI Sites with a Three-County Empirical Distribution for Each Variable

Indicator	Toxic Site Mean	Control Site Mean
Population	8,028.5	4,617.5
Population density	2,573.2	1,480.0
Families	1,765.3	1,172.9
Households	3,583.2	1,847.4
White	6,180.5	3,896.4
Black	714.4	263.3
Asian	872.5	347.3
Indian	131.2	55.8
Hispanic	295.2	138.8
% white	78.2	85.4
% Asian	11.0	7.6
% black	9.0	5.8
% Indian	1.7	1.2
% minority	20.0	8.5
Households median income	$29,155	$40,429
Average per capita income	$16,198	$17,160
Households with earnings	2,876.4	1,550.7
Households without earnings	706.7	296.8
% households without earnings	19.7	16.1
% households with income <$15K	31.2	22.5
% households with income >$100K	3.4	4.5
Housing units	3,829.4	1,941.3
% housing units owner occupied	42.8	57.6
% without high school diploma	16.2	13.6
% college grad[a]	29.6	27.9
% children in poverty	17.8	10.3
% children age 5 or under	7.7	8.8
% age 65 or older	12.7	10.7
% households w/o earnings	19.7	16.0
% persons in poverty	13.9	9.0
% age 5 or younger in poverty	20.8	13.5
% age 65 older in poverty	10.5	7.8

[a] Difference between values for this variable not statistically significant.

nearest neighbors, as well as the estimated sample variance of the state or nearest neighbor weights, shrunk the observed value toward the center of the distribution.

A comparison of observed versus spatial Bayesian SMR is illustrated in Figure 3. Comparing the two maps, the map of observed SMRs shows how the SMR is lowered; darker blue shades in the thematic map indicate an SMR below 100, while those above 100 tend toward darker shades of red.

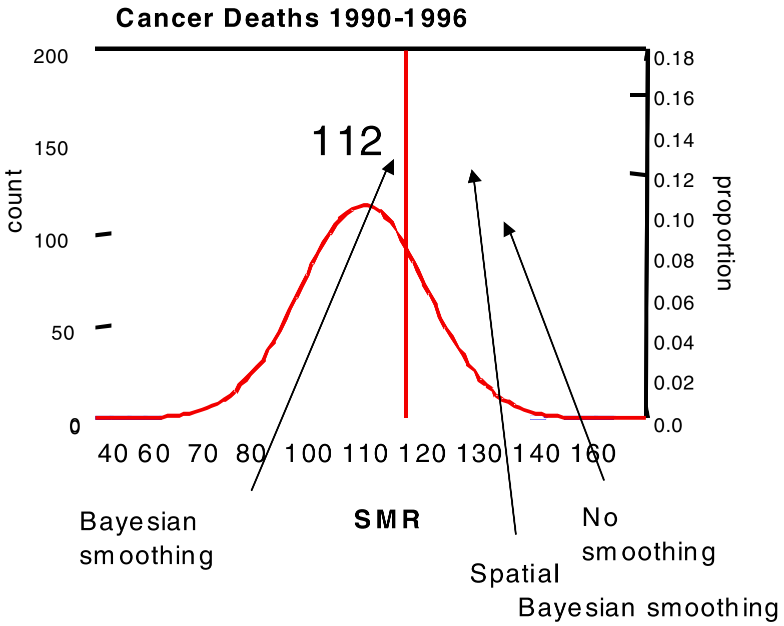

Figure 2 Standard mortality ratio (SMR) for observed, Bayesian and spatial Bayesian smoothing.

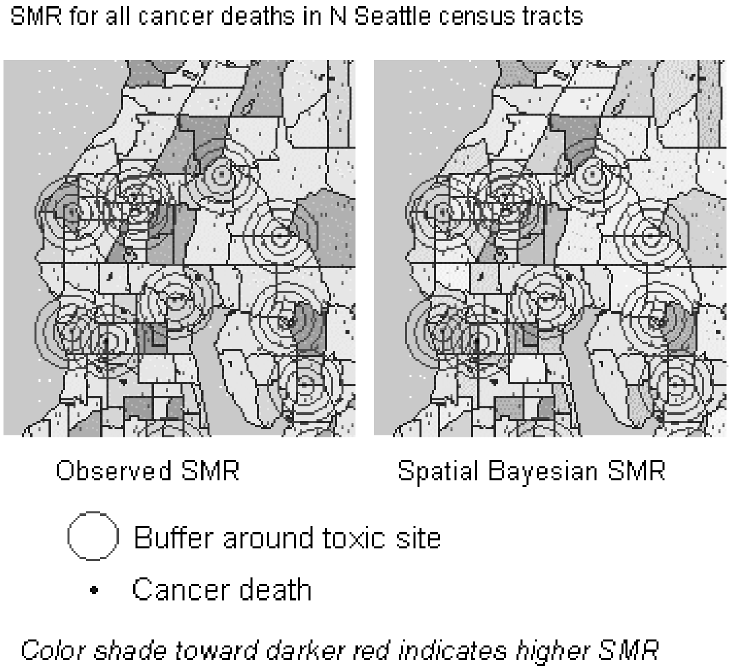
SMR for all cancer deaths in N Seattle census tracts
Observed SMR
Spatial Bayesian SMR
Buffer around toxic site
Cancer death
Color shade toward darker red indicates higher SMR

Figure 3 Thematic maps of cancer SMR before and after Bayesian smoothing.

Discussion

We have presented some initial results about a methodology for determining the characteristics of populations around SF/TRI sites. The Monte Carlo simulation builds an empirical distribution that one can use to compare a profile of one or more toxic sites with the rest of the county or region. One advantage of this method is that it requires no assumption about the shape of the distribution. It may also account for some spatial autocorrelation because the random selection of the control points is not dependent on the location of other points selected in the set. The method is simple to use and may give better results than ordinary parametric methods.

In the results presented here, it is clear that populations around the toxic sites identified from the Landview III database were more likely to be minority and lower income. In this preliminary work, however, no distinction was made between sites with respect to harmfulness or to content of known carcinogenic compounds or other putative substances known to affect health. Subsequent studies will look at neighborhoods around sites to consider what the site contains and, perhaps most importantly, its remediation status.

A similar scheme was used to assess disease rates (via death rates) around the toxic sites. There appeared to be some elevation of the death rate for all cancer deaths, but after Bayesian smoothing the differences were not important. No adjustment was made to account for the increase in death rates that is often associated with minority or lower-income populations. This might be done by calculating the SMR with race and sex stratification, as well as with the usual 5- or 10-year age groups.

As with all Monte Carlo based schemes, the quality of the results depends on the sampling protocol for the control points. Future studies need to consider adjusting the probability of selecting a particular control point based on the historical land use designation. Zoning laws that were in effect many years ago certainly influenced whether or not a potential SF/TRI site could ever be built at a particular location. In the current model, all points are equally likely to be sampled, presuming they are not under water or located in places such as cemeteries and older parks.

References

1. US Department of Commerce, US Environmental Protection Agency. 1997. *Landview III environmental mapping software.* www.census.gov/apsd/pp98/pp.html.
2. GeoLytics, Inc. 1998. *Census CD + maps, US 1990 Census (STF3A, C, and D).* East Brunswick, NJ: GeoLytics, Inc. www.Geolytics.com.
3. Claritas Corporation. 1990–1997. *Annual census data.* Arlington, VA: Claritas Corporation.
4. Caliper Corporation. 1998. *Maptitude, v 4.02.* Newton, MA: Caliper Corporation. www.caliper.com.
5. Bailey TC, Gatrell AC. 1995. Empirical Bayes estimation. In: *Interactive spatial data analysis.* Essex, England: Longman Scientific & Technical, Longman Group, Ltd. 303–98.
6. Devine OJ, Louis TA. 1994. A constrained empirical Bayes estimator for incidence rates in areas with small populations. *Statistics in Medicine* 13(11):1119–33.
7. Devine OJ, Louis TA, Halloran ME. 1994. Empirical Bayes methods for stabilizing incidence rates before mapping. *Epidemiology* 5(6):622–30.